A Consistent Likelihood-Based Variable Selection Method in Normal Multivariate Linear Regression

Ryoya Oda[*] [†] and Hirokazu Yanagihara[‡]

(Last Modified: December 18, 2020)

Abstract

We propose a likelihood-based variable selection method for normality-assumed multivariate linear regression contexts. The method is reasonably fast, and it exhibits adequate selection consistency, when the sample size always tends to infinity, but the numbers of response and explanatory variables do not necessarily have to tend to infinity. It can be expected that the proposed method has a high probability of selecting a true subset under a moderate sample size.

1 Introduction

Multivariate linear regression with an $n \times p$ response matrix $\boldsymbol{Y}$ and an $n \times k$ explanatory matrix $\boldsymbol{X}$ is one of the fundamental methods of inferential statistical analysis and it is introduced in many statistical textbooks (e.g., [10, 12]), where n is the sample size, and p and k are the numbers of response variables and explanatory variables, respectively. Note $N = n-p-k+1$. We assume that $N - 2 > 0$ and $\mathrm{rank}(\boldsymbol{X}) = k < n$. Let $\omega = \{1, \ldots, k\}$ be the full set consisting of all the column indexes of $\boldsymbol{X}$, and let $\boldsymbol{X}_j$ be an $n \times k_j$ matrix consisting of columns of $\boldsymbol{X}$ indexed by the elements of $j \subset \omega$, where k_j is the number of elements in j, i.e., $k_j = \#(j)$. From the above notation, it holds that $\boldsymbol{X} = \boldsymbol{X}_\omega$. For a subset $j \subset \omega$, the multivariate linear regression model with $\boldsymbol{Y}$ and $\boldsymbol{X}_j$ is expressed as follows:

$$\boldsymbol{Y} \sim \mathcal{N}_{n\times p}(\boldsymbol{X}_j \boldsymbol{\Theta}_j, \boldsymbol{\Sigma}_j \otimes \boldsymbol{I}_n), \tag{1}$$

where $\boldsymbol{\Theta}_j$ is a $k_j \times p$ matrix of regression coefficients and $\boldsymbol{\Sigma}_j$ is a $p \times p$ covariance matrix. In actual empirical contexts, it is important to examine which of the k explanatory variables affect the response variables, and this is regarded as the problem of selecting a best model from (1), in other words, selecting a best subset of ω. To achieve this, it is common to search over all candidate subsets and a variable selection criterion (SC) is often used to identify an optimal model following this search. Akaike's information criterion (AIC) [1, 2] is the most widely applied

[*] Corresponding author. Email: ryoya-oda@hiroshima-u.ac.jp

[†] School of Informatics and Data Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan

[‡] Graduate School of Advanced Science and Engineering, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima, Japan

tool in this respect. For $j \subset \omega$, let $\hat{\boldsymbol{\Sigma}}_j$ be the maximum likelihood estimator of $\boldsymbol{\Sigma}_j$ in (1), which is defined by

$$\hat{\boldsymbol{\Sigma}}_j = n^{-1}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_j)\boldsymbol{Y},$$

where $\boldsymbol{P}_j$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{X}_j$, i.e., $\boldsymbol{P}_j = \boldsymbol{X}_j(\boldsymbol{X}_j'\boldsymbol{X}_j)^{-1}\boldsymbol{X}_j'$. A generalized information criterion (GIC) [7] for $j \subset \omega$ is defined by

$$\mathrm{GIC}(j) = n\log|\hat{\boldsymbol{\Sigma}}_j| + np(\log 2\pi + 1) + \alpha\left\{pk_j + \frac{p(p+1)}{2}\right\},$$

where α is a positive constant and means the strength of the penalty term for j. Specifying the value of α, the GIC includes several criteria as special cases, e.g., the AIC ($\alpha = 2$), the Bayesian information criterion (BIC) [11] ($\alpha = \log n$) and the Hannan-Quinn information criterion (HQC) [5] ($\alpha = 2\log\log n$).

Recently, there has been significant attention in the literature to statistical methods for high-dimensional data. In high-dimensional data contexts in which the number of explanatory variables is substantial, it may take an inordinate amount of time to search for and identify the best subset by calculating variable selection criteria for all the candidate subsets. For practical reasons, we focus on a selection method based on an SC. Let ω_ℓ be the complement set of $\{\ell\}$ for ω, i.e., $\omega_\ell = \omega\backslash\{\ell\}$. Then, the following best subset according to the selection method based on an SC is presented:

$$\{\ell \in \omega \mid \mathrm{SC}(\omega_\ell) > \mathrm{SC}(\omega)\}, \tag{2}$$

where $\mathrm{SC}(j)$ is the value of an SC for $j \subset \omega$. Method (2) is introduced by [13] and is referred to as the kick-one-out method by [3]. In particular, method (2) based on the GIC is expressed as

$$\hat{j} = \{\ell \in \omega \mid \mathrm{GIC}(\omega_\ell) > \mathrm{GIC}(\omega)\} = \{\ell \in \omega \mid n\log|\hat{\boldsymbol{\Sigma}}_{\omega_\ell}\hat{\boldsymbol{\Sigma}}_\omega^{-1}| > p\alpha\}. \tag{3}$$

Method (3) can be regarded as a likelihood-based selection method. In multivariate linear regression contexts, method (3) are also used by [3, 9], and method (2) based on the generalized C_p (GC_p) criterion [6] is used by [3, 8].

In this paper, we focus on the likelihood-based selection method (3). We assume that the data are generated from the following true model for the true subset j_*:

$$\boldsymbol{Y} \sim \mathcal{N}_{n\times p}(\boldsymbol{X}_{j_*}\boldsymbol{\Theta}_*, \boldsymbol{\Sigma}_* \otimes \boldsymbol{I}_n), \tag{4}$$

where $\boldsymbol{\Theta}_*$ is a $k_{j_*} \times p$ matrix of true regression coefficients wherein the row vectors are not zeros and $\boldsymbol{\Sigma}_*$ is a $p \times p$ true covariance matrix which is positive definite. We state that method (3) has selection consistency if $P(\hat{j} = j_*) \to 1$ holds. The purpose of this paper is to propose a variable selection method using (3), which will be reasonably fast and meet the conditions for the selection consistency $P(\hat{j} = j_*) \to 1$ under the following high-dimensional (HD) asymptotic framework:

$$\mathrm{HD}: n \to \infty,\ \frac{p+k}{n} \to c \in [0, 1). \tag{5}$$

The HD asymptotic framework means that n always tends to infinity, but p, k and k_{j_*} do not necessarily have to tend to infinity. Hence, it is expected that the proposed method has a high probability of selecting the true subset j_* under moderate sample sizes even when p, k and k_{j_*} are large. As related researches in high-dimensional contexts, [8] examined the selection consistency of (2) based on the GC_p criterion under the HD asymptotic framework. When the true model is generated by a non-normal distribution, [3] also examined a strong selection consistency of (3) and (2) based on the GC_p criterion under the HD asymptotic framework with the exception that $k_{j_*} \to \infty$ and $c = 0$. Note that the strong selection consistency means that $P(\hat{j} \to j_*) = 1$ holds and is stronger than the selection consistency $P(\hat{j} = j_*) \to 1$. In this paper, the conditions for selection consistency $P(\hat{j} = j_*) \to 1$ under the HD asymptotic framework are derived on the basis of the method by [8].

The remainder of the paper is organized as follows. In section 2, we obtain the conditions for selection consistency of (3) and propose a consistent selection method by the obtained conditions under the HD asymptotic framework. In section 3, we conduct numerical experiments for verification purposes. Technical details are relegated to the Appendix.

2 Proposed Selection Method

First, we prepare notation and assumptions for conditions for the selection consistency of (3) $P(\hat{j} = j_*) \to 1$ under the HD asymptotic framework. To obtain the conditions for the selection consistency of (3), the following three assumptions are prepared:

Assumption A1 *The true subset j_* is included in the full set ω, i.e., $j_* \subset \omega$.*

Assumption A2 *There exists $c_1 > 0$ such that*

$$n^{-1} \min_{\ell \in j_*} \boldsymbol{x}_\ell' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{x}_\ell \geq c_1,$$

where $\boldsymbol{x}_\ell$ is the ℓ-th column vector of $\boldsymbol{X}$.

Assumption A3 *There exist $c_2 > 0$ and $c_3 \geq 1/2$ such that*

$$n^{1-c_3} \min_{\ell \in j_*} \boldsymbol{\theta}_\ell' \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_\ell \geq c_2,$$

where $\boldsymbol{\theta}_\ell$ is the ℓ-th row vector of $\boldsymbol{\Theta}_*$.

Assumption A1 is needed to consider the selection consistency. Assumption A2 and Assumption A3 concern the asymptotic restriction for explanatory variables and parameters for the true model. These assumptions are also used by [8] and they allow the minimum eigenvalue of $n^{-1}\boldsymbol{X}'\boldsymbol{X}$ to be bounded away from 0 and the minimum value of $\boldsymbol{\theta}_\ell' \boldsymbol{\Sigma}_*^{-1} \boldsymbol{\theta}_\ell$ to vanish to 0 at a slow speed. For $\ell \in \omega$, let a $p \times p$ non-centrality matrix and parameter be denoted by

$$\boldsymbol{\Delta}_\ell = \boldsymbol{\Sigma}_*^{-1/2} \boldsymbol{\Theta}_*' \boldsymbol{X}_{j_*}' (\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell}) \boldsymbol{X}_{j_*} \boldsymbol{\Theta}_* \boldsymbol{\Sigma}_*^{-1/2}, \ \delta_\ell = \mathrm{tr}(\boldsymbol{\Delta}_\ell). \tag{6}$$

It should be emphasized that $\boldsymbol{\Delta}_\ell = \boldsymbol{O}_{p,p}$ and $\delta_\ell = 0$ hold if and only if $\ell \notin j_*$ under Assumption A1, where $\boldsymbol{O}_{p,p}$ is a $p \times p$ matrix of zeros.

Next, we obtain the conditions for the selection consistency of (3). The following lemma is prepared to examine the distribution of $|\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}|$ (the proof is given in Appendix 1):

Lemma 1 *For $\ell \in \omega$, let u_ℓ and v_ℓ be independent random variables distributed according to $u_\ell \sim \chi^2(p;\delta_\ell)$ and $v_\ell \sim \chi^2(N)$, respectively, where δ_ℓ is defined by* (6)*. Then, under Assumption A1, we have*

$$|\hat{\mathbf{\Sigma}}_{\omega_\ell}\hat{\mathbf{\Sigma}}_\omega^{-1}| = 1 + \frac{u_\ell}{v_\ell}.$$

By using Lemma 1, the conditions for the selection consistency of (3) are obtained in Theorem 1 (the proof is given in Appendix 2).

Theorem 1 *Suppose that Assumptions A1, A2 and A3 hold. Then, the selection method $\hat{j}$ exhibits selection consistency under the HD asymptotic framework* (5)*, i.e., $P(\hat{j} = j_*) \to 1$ holds, if for some $r \in \mathbb{N}$ the following conditions are satisfied:*

$$\alpha = \frac{n}{p}\log\left(1 + \frac{p}{N-2} + \beta\frac{p}{N-2}\right), \ \beta > 0, \ \text{s.t.} \ \frac{\sqrt{p}}{k^{1/2r}}\beta \to \infty, \ \frac{p}{n^{c_3}}\beta \to 0. \tag{7}$$

By using the result of Theorem 1, we propose the consistent likelihood-based selection method $\hat{j}$ with the following value of α:

$$\alpha = \tilde{\alpha} = \frac{n}{p}\log\left(1 + \frac{p}{N-2} + \frac{k^{1/4}\sqrt{p}\log n}{N-2}\right). \tag{8}$$

From (7), it is straightforward that the selection method $\hat{j}$ with $\alpha = \tilde{\alpha}$ has selection consistency under the HD asymptotic framework when $c_3 > 3/4$.

Remark 1 (Relationship with the GC_p criterion) *The difference between the GC_p criterion for ω_ℓ ($\ell \in \omega$) and that for ω is defined as*

$$GC_p(\omega_\ell) - GC_p(\omega) = (n-k)\mathrm{tr}(\hat{\mathbf{\Sigma}}_{\omega_\ell}\hat{\mathbf{\Sigma}}_\omega^{-1}) - p\gamma,$$

where γ is a positive constant. From Lemma A.1 in [8], it is known that the equation $\mathrm{tr}(\hat{\mathbf{\Sigma}}_{\omega_\ell}\hat{\mathbf{\Sigma}}_\omega^{-1}) = u_\ell v_\ell^{-1}$ holds, where u_ℓ and v_ℓ are defined in Lemma 1. This fact implies that the likelihood-based selection method (3) *can be regarded as equivalent to* (2) *based on the GC_p criterion when α and γ are adjusted adequately. Especially, the proposed method* (3) *with $\alpha = \tilde{\alpha}$ can be regarded as nearly identical to the method in [8].*

Finally, we present an efficient calculation for high-dimensional data. When p and k are large, $|\hat{\mathbf{\Sigma}}_{\omega_\ell}\hat{\mathbf{\Sigma}}_\omega^{-1}|$ in (3) should not be calculated simply, because the size of $\hat{\mathbf{\Sigma}}_{\omega_\ell}$ is $p \times p$ and moreover $\boldsymbol{P}_{\omega_\ell}$ must be calculated to derive $\hat{\mathbf{\Sigma}}_{\omega_\ell}$ for each $\ell \in \omega$. For $\ell \in \omega$, let r_ℓ be the (ℓ,ℓ)-th element of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and let $\boldsymbol{z}_\ell$ be the ℓ-th column vector of $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. In accordance with [8], it holds that $\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_\ell} = r_\ell^{-1}\boldsymbol{z}_\ell\boldsymbol{z}_\ell'$. Hence, it is straightforward that $|\hat{\mathbf{\Sigma}}_{\omega_\ell}\hat{\mathbf{\Sigma}}_\omega^{-1}|$ can be expressed as

$$|\hat{\mathbf{\Sigma}}_{\omega_\ell}\hat{\mathbf{\Sigma}}_\omega^{-1}| = 1 + n^{-1}r_\ell^{-1}\boldsymbol{z}_\ell'\boldsymbol{Y}\hat{\mathbf{\Sigma}}_\omega^{-1}\boldsymbol{Y}'\boldsymbol{z}_\ell. \tag{9}$$

Therefore, we recommend calculating $\hat{j}$ using (9).

3 Numerical Studies

We present numerical results to compare the probabilities of selecting the true subset j_* by the proposed method (3) with $\alpha = \tilde{\alpha}$ in (8) and the two methods (3) based on the AIC and

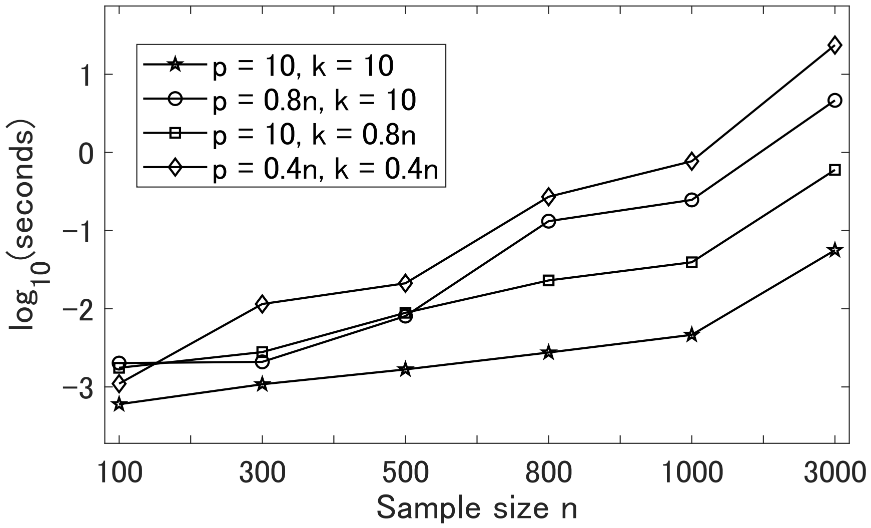

Figure 1: Base-10 logarithm of CPU times associated with executing the proposed method (3) with $\alpha = \tilde{\alpha}$ in (8)

BIC ($\alpha = 2, \log n$). Moreover, we present the CPU times associated with executing the proposed method (3) with $\alpha = \tilde{\alpha}$ in (8). The probabilities and CPU times were calculated by Monte Carlo simulations with $10,000$ iterations executed in MATLAB 9.6.0 on an Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz 3601 Mhz, 8 cores, 16 logical processors and 64 GB of RAM. We set the true subset and the number of the true explanatory variables as $j_* = \{1, \ldots, k_{j_*}\}$ and $k_{j_*} = k/2$. The data $\boldsymbol{Y}$ were generated by the true model (4) and $\boldsymbol{X}$ and the true parameters were determined as follows:

$$\boldsymbol{X} \sim \mathcal{N}_{n\times k}(\boldsymbol{O}_{n,k}, \boldsymbol{\Psi} \otimes \boldsymbol{I}_n),\ \boldsymbol{\Theta}_* = \boldsymbol{1}_{k_{j_*}}\boldsymbol{1}_p',\ \boldsymbol{\Sigma}_* = 0.4\{(1-0.8)\boldsymbol{I}_p + 0.8\boldsymbol{1}_p\boldsymbol{1}_p'\},$$

where the (a, b)-th element of $\boldsymbol{\Psi}$ is $(0.5)^{|a-b|}$ and $\boldsymbol{1}_p$ is a p-dimensional vector of ones.

Table 1 shows the selection probabilities. Therein, j_- and j_+ denote the underspecified and overspecified subsets of ω satisfying $j_- \cap j_* = \emptyset$ and $j_+ \supsetneq j_*$, respectively. From Table 1, we observe that the proposed method appears to have the selection consistency $P(\hat{j} = j_*) \to 1$ under the HD asymptotic framework. However, the method (3) based on the AIC tends towards selecting an overspecified subsets j_+. The method (3) based on the BIC seems to exhibit the selection consistency $P(\hat{j} = j_*) \to 1$ when only n tends to infinity.

Figure 1 shows the base-10 logarithm of CPU times by the proposed method (3) with $\alpha = \tilde{\alpha}$ in (8). From Figure 1, we observe that the proposed method is fast (about 24 seconds at its slowest, i.e., when $n = 3000, p = 1200, k = 1200$).

Appendix 1: Proof of Lemma 1

This proof is based on that of Lemma A.1 in [8]. For $\ell \in \omega$, let $\boldsymbol{W}_\ell = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{P}_\omega - \boldsymbol{P}_{\omega_\ell})\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$ and $\boldsymbol{W} = \boldsymbol{\Sigma}_*^{-1/2}\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{P}_\omega)\boldsymbol{Y}\boldsymbol{\Sigma}_*^{-1/2}$. Then, from a property of the non-central

Table 1: Selection probabilities (%) of j_-, j_* and j_+ by the proposed method (3) with $\alpha = \tilde{\alpha}$ in (8) and the two methods (3) based on the AIC and BIC

			Proposed			AIC			BIC		
n	p	k	j_-	j_*	j_+	j_-	j_*	j_+	j_-	j_*	j_+
100	10	10	0.00	99.84	0.16	0.00	70.05	29.95	0.00	99.97	0.03
300	10	10	0.00	99.98	0.02	0.00	82.10	17.90	0.00	100.00	0.00
500	10	10	0.00	100.00	0.00	0.00	83.66	16.34	0.00	100.00	0.00
800	10	10	0.00	100.00	0.00	0.00	84.94	15.06	0.00	100.00	0.00
1000	10	10	0.00	100.00	0.00	0.00	85.76	14.24	0.00	100.00	0.00
3000	10	10	0.00	100.00	0.00	0.00	86.13	13.87	0.00	100.00	0.00
100	80	10	58.90	29.26	11.84	0.00	0.00	100.00	98.27	1.67	0.06
300	240	10	0.34	94.87	4.79	0.00	0.12	99.88	100.00	0.00	0.00
500	400	10	0.00	97.68	2.32	0.00	0.29	99.71	100.00	0.00	0.00
800	640	10	0.00	99.21	0.79	0.00	0.50	99.50	100.00	0.00	0.00
1000	800	10	0.00	99.42	0.58	0.00	0.60	99.40	100.00	0.00	0.00
3000	2400	10	0.00	99.96	0.04	0.00	0.91	99.09	100.00	0.00	0.00
100	10	80	99.85	0.11	0.04	0.00	0.00	100.00	1.31	0.00	98.69
300	10	240	96.02	3.98	0.00	0.00	0.00	100.00	0.00	0.00	100.00
500	10	400	42.73	57.27	0.00	0.00	0.00	100.00	0.00	0.00	100.00
800	10	640	0.02	99.98	0.00	0.00	0.00	100.00	0.00	0.00	100.00
1000	10	800	0.02	99.98	0.00	0.00	0.00	100.00	0.00	0.00	100.00
3000	10	2400	0.00	100.00	0.00	0.00	0.00	100.00	0.00	0.00	100.00
100	40	40	90.19	8.88	0.93	0.00	0.00	100.00	81.44	15.72	2.84
300	120	120	80.70	19.25	0.05	0.00	0.00	100.00	100.00	0.00	0.00
500	200	200	49.99	50.01	0.00	0.00	0.00	100.00	100.00	0.00	0.00
800	320	320	12.10	87.90	0.00	0.00	0.00	100.00	100.00	0.00	0.00
1000	400	400	4.48	95.52	0.00	0.00	0.00	100.00	100.00	0.00	0.00
3000	1200	1200	0.00	100.00	0.00	0.00	0.00	100.00	100.00	0.00	0.00

Wishart distribution and Cochran's Theorem (e.g., [4]), we can see that $\boldsymbol{W}_\ell$ and $\boldsymbol{W}$ are independent and $\boldsymbol{W}_\ell \sim \mathcal{W}_p(1, \boldsymbol{I}_p; \boldsymbol{\Delta}_\ell)$ and $\boldsymbol{W} \sim \mathcal{W}_p(n-k, \boldsymbol{I}_p)$, where $\boldsymbol{\Delta}_\ell$ is defined by (6). Since the rank of $\boldsymbol{\Delta}_\ell$ is 1, we decompose $\boldsymbol{\Delta}_\ell$ as $\boldsymbol{\Delta}_\ell = \boldsymbol{\eta}_\ell \boldsymbol{\eta}_\ell'$ by using a p-dimensional vector $\boldsymbol{\eta}_\ell$. By using $\boldsymbol{\eta}_\ell$, we can express $\boldsymbol{W}_\ell$ as $\boldsymbol{W}_\ell = (\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)(\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)'$, where $\boldsymbol{\varepsilon}_\ell \sim \mathcal{N}_{p\times 1}(\boldsymbol{0}_p, 1 \otimes \boldsymbol{I}_p)$ and $\boldsymbol{\varepsilon}_\ell$ is independent of $\boldsymbol{W}$. Then, we have

$$|\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}| = |\boldsymbol{I}_p + \boldsymbol{W}_\ell \boldsymbol{W}^{-1}| = 1 + (\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)' \boldsymbol{W}^{-1} (\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell). \tag{10}$$

Let u_ℓ and v_ℓ be constants as follows:

$$u_\ell = (\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)'(\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell), \ v_\ell = \frac{(\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)'(\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)}{(\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)' \boldsymbol{W}^{-1} (\boldsymbol{\varepsilon}_\ell + \boldsymbol{\eta}_\ell)}.$$

Then, from a property of the Wishart distribution, we can state that u_ℓ and v_ℓ are independent, $u_\ell \sim \chi^2(p; \delta_\ell)$ and $v_\ell \sim \chi^2(N)$. Using (10), $|\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}|$ is expressed as

$$|\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}| = 1 + \frac{u_\ell}{v_\ell}.$$

Therefore, the proof of Lemma 1 is completed. □

Appendix 2: Proof of Theorem 1

To prove Theorem 1, we use the the following results about the divergence orders of central moments, which is seen in [8, Lemma A.2].

Lemma 1 ([8]) *Let δ be a positive constant. And let t_1, t_2 and v be random variables distributed according to $\chi^2(p)$, $\chi^2(p; \delta)$ and $\chi^2(N)$, respectively, where t_1 and t_2 are independent of v. Then, under the HD asymptotic framework* (5)*, for $N - 4r > 0$ ($r \in \mathbb{N}$), we have*

$$E\left[\left(\frac{t_1}{v} - \frac{p}{N-2}\right)^{2r}\right] = O(p^r n^{-2r}),$$

$$E\left[\left(\frac{t_2}{v} - \frac{p+\delta}{N-2}\right)^{2r}\right] = O(\max\{(p+\delta)^r n^{-2r}, (p+\delta)^{2r} n^{-3r}\}).$$

From Lemma 1, it holds that $|\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}| = 1 + u_\ell v_\ell^{-1}$, where u_ℓ and v_ℓ are independent random variables distributed according to $u_\ell \sim \chi^2(p; \delta_\ell)$ and $v_\ell \sim \chi^2(N)$, respectively. The upper bound of $P(\hat{j} \neq j_*)$ is expressed as

$$P(\hat{j} \neq j_*) \leq \sum_{\ell \notin j_*} P(n \log |\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}| \geq p\alpha) + \sum_{\ell \in j_*} P(n \log |\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}| \leq p\alpha).$$

Hence, we should prove that the right-hand side of the above inequality tends to 0 under the HD asymptotic framework. First, we consider the case of $\ell \notin j_*$. Note that $\boldsymbol{\Delta}_\ell = \boldsymbol{O}_{p,p}$, i.e., $\delta_\ell = 0$ for $\ell \notin j_*$. Let $\rho_\beta = p\beta(N-2)^{-1}$. Then, using Markov's inequality and (7), for any $r \in \mathbb{N}$, we have

$$\begin{aligned} &\sum_{\ell \notin j_*} P(n \log |\hat{\boldsymbol{\Sigma}}_{\omega_\ell} \hat{\boldsymbol{\Sigma}}_\omega^{-1}| \geq p\alpha) = \sum_{\ell \notin j_*} P\left(\frac{u_\ell}{v_\ell} - \frac{p}{N-2} \geq \rho_\beta\right) \\ &\leq (k - k_{j_*}) P\left(\left|\frac{u_1}{v_1} - \frac{p}{N-2}\right| \geq \rho_\beta\right) \leq k \rho_\beta^{-2r} E\left[\left(\frac{u_1}{v_1} - \frac{p}{N-2}\right)^{2r}\right]. \end{aligned}$$

From Lemma 1, the moment in the above inequality has $O(p^r n^{-2r})$. Hence, we have

$$\sum_{\ell \notin j_*} P(n \log |\hat{\mathbf{\Sigma}}_{\omega_\ell} \hat{\mathbf{\Sigma}}_\omega^{-1}| \geq p\alpha) = O(kp^{-r}\beta^{-2r}) = o(1). \tag{11}$$

Next, we consider the case of $\ell \in j_*$. Note that $\delta_\ell > 0$ for $\ell \in j_*$. Let $\delta_{\min} = \min_{\ell \in j_*} \delta_\ell$. Then, from Assumption A2 and Assumption A3, the inequality $n^{-c_3}\delta_{\min} \geq c_1 c_2$ holds because of $\delta_{\min} \geq \min_{\ell \in j_*} \boldsymbol{x}_\ell'(\boldsymbol{I}_n - \boldsymbol{P}_{\omega_\ell})\boldsymbol{x}_\ell \boldsymbol{\theta}_\ell' \mathbf{\Sigma}_*^{-1} \boldsymbol{\theta}_\ell$. Hence, we have $\rho_\beta \delta_{\min}^{-1} = o(1)$ from (7). Then, using Markov's inequality, for sufficiently large N and any $r \in \mathbb{N}$, we have

$$\begin{aligned}
\sum_{\ell \in j_*} P(n \log |\hat{\mathbf{\Sigma}}_{\omega_\ell} \hat{\mathbf{\Sigma}}_\omega^{-1}| \leq p\alpha) &= \sum_{\ell \in j_*} P\left(\frac{u_\ell}{v_\ell} - \frac{p + \delta_\ell}{N-2} \leq \rho_\beta - \frac{\delta_\ell}{N-2} \right) \\
&\leq \sum_{\ell \in j_*} P\left(\left| \frac{u_\ell}{v_\ell} - \frac{p + \delta_\ell}{N-2} \right| \geq -\rho_\beta + \frac{\delta_\ell}{N-2} \right) \\
&\leq k_{j_*} \max_{\ell \in j_*} \left(-\rho_\beta + \frac{\delta_\ell}{N-2} \right)^{-2r} E\left[\left(\frac{u_\ell}{v_\ell} - \frac{p+\delta_\ell}{N-2} \right)^{2r} \right].
\end{aligned} \tag{12}$$

From Lemma 1, the maximum value except for constant parts of the above moment is $(p + \delta_\ell)^r n^{-2r}$ or $(p + \delta_\ell)^{2r} n^{-3r}$. Hence, for sufficiently large $r \in \mathbb{N}$, we have

$$\begin{aligned}
& k_{j_*} \max_{\ell \in j_*} \left(-\rho_\beta + \frac{\delta_\ell}{N-2} \right)^{-2r} \{ (p + \delta_\ell)^r n^{-2r} + (p + \delta_\ell)^{2r} n^{-3r} \} \\
&\leq k_{j_*} \max_{\ell \in j_*} \left(1 - \rho_\beta \frac{N-2}{\delta_\ell} \right)^{-2r} \left\{ \left(1 + p\delta_{\min}^{-1}\right)^r \delta_{\min}^{-r} + \left(1 + p\delta_{\min}^{-1}\right)^{2r} n^{-r} \right\} \\
&= O(k_{j_*} \delta_{\min}^{-r}) + O(k_{j_*} p^r \delta_{\min}^{-2r}) + O(k_{j_*} n^{-r}) + O(k_{j_*} p^{2r} n^{-r} \delta_{\min}^{-2r}).
\end{aligned} \tag{13}$$

We can see that (13) tends to 0 under the HD asymptotic framework if $c_3 > (1 + r^{-1})/2$. Since r is arbitrary, the inequality $c_3 > (1 + r^{-1})/2$ is equivalent to $c_3 \geq 1/2$. Hence, from (12) and (13), we have

$$\sum_{\ell \in j_*} P(n \log |\hat{\mathbf{\Sigma}}_{\omega_\ell} \hat{\mathbf{\Sigma}}_\omega^{-1}| \leq p\alpha) = o(1). \tag{14}$$

Therefore, (11) and (14) completes the proof of Theorem 1. □

Acknowledgments

This work was supported by funding from JSPS KAKENHI (grant numbers JP20K14363, JP20H04151, and JP19K21672 to Ryoya Oda; and JP16H03606, JP18K03415, and JP20H04151 to Hirokazu Yanagihara).

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (eds. B. N. Petrov & F. Csáki), 995–1010. Akadémiai Kiadó, Budapest. doi:10.1007/978-1-4612-1694-0_15

[2] Akaike, H. (1974). A new look at the statistical model identification. *Institute of Electrical and Electronics Engineers. Transactions on Automatic Control* **AC-19** 716–723. doi:10.1109/TAC.1974.1100705

[3] Bai, Z. D., Fujikoshi, Y. and Hu, J. (2018). Strong consistency of the AIC, BIC, C_p and KOO methods in high-dimensional multivariate linear regression. TR No. 18–9, *Statistical Research Group*, Hiroshima University.

[4] Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations.* John Wiley & Sons, Inc., Hoboken, New Jersey.

[5] Hannan, E. J. and Quinn, B. G (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **26** 270–273. doi:10.1111/j.2517-6161.1979.tb01072.x

[6] Nagai, I., Yanagihara, H. and Satoh, K. (2012). Optimization of ridge parameters in multivariate generalized ridge regression by plug-in methods. *Hiroshima Math. J.* **42** 301–324. doi:10.32917/hmj/1355238371

[7] Nishii, R., Bai, Z. D. and Krishnaiah, P. R. (1988). Strong consistency of the information criterion for model selection in multivariate analysis. *Hiroshima Math. J.* **18** 451–462. doi:10.32917/hmj/1206129611

[8] Oda, R. and Yanagihara, H. (2020). A fast and consistent variable selection method for high-dimensional multivariate linear regression with a large number of explanatory variables. *Electron. J. Statist.* **14** 1386-1412. doi:10.1214/20-EJS1701

[9] Sakurai, T. and Fujikoshi, Y. (2017). High-dimensional properties of information criteria and their efficient criteria for multivariate linear regression models with covariance structures. TR No. 17–13, *Statistical Research Group*, Hiroshima University.

[10] Srivastava, M. S. (2002). *Methods of Multivariate Statistics.* John Wiley & Sons, New York.

[11] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. doi:10.1214/aos/1176344136

[12] Timm, N. H. (2002). *Applied Multivariate Analysis.* Springer-Verlag, New York.

[13] Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* **20** 1–25. doi:10.1016/0047-259X(86)90017-5